

Linear Regression Practice: Movies & Bechdel Test

Setup

Assign roles in your group:

1. Manager. Have instructions available for reference and read aloud as needed. Keep group on task. Try to budget time to complete tasks. Make sure everyone is included in the process, asking for input from those who keep quiet.
2. R Typist. Main editor of your group qmd file. Creates project, types code and text in RStudio.
3. Graphics Reference. Have graphics tutorials open for code reference and ideas.
4. Class notes reference. Have notes from class available for code and concepts from class.
5. Writer and Format Police. Help with phrasing for written answers. Make sure the qmd file has titled sections and watch for typos in text and code.

If you have 4 team members, combine roles 4-5. If you have 5, divide 4 into writing and format roles.

Data Preparation

All code needed for data prep is provided for you – just copy and run it. You should try to get a rough sense for what each bit of code is doing, but you don't need to be able to do things like this on your own.

The dataset is available in the `fivethirtyeight` R package, and is called `bechdel`. It is discussed in a FiveThirtyEight.com [article about women in Hollywood](#). It is a play on the Bechdel Test, a measure of women's roles in a film based on a 1985 comic by Alison Bechdel:



(of course, this is *not* a perfect test.) To load the dataset and read details about it, run:

```
library(fivethirtyeight)
?bechdel
```

Before analysis, we will limit the data to the years 1990 - 2013, since the dataset creators say “the data has significantly more depth since then”. The code below (given to you) will do the job:

```
bechdel_13 <- bechdel |>
  filter(year >= 1990 & year <= 2013)
```

From now on, we will use the “since 2013” dataset, `bechdel_13`, for all analysis.

We can also add a few additional variables to the data set.

- `roi`: Total return on investment (gross earnings divided by total budget)
- `profit`: International gross minus budget

```
bechdel_13 <- bechdel_13 |>
  mutate(roi = intgross_2013 / budget_2013,
        profit = intgross_2013 - budget_2013)
```

Finally, we will adjust the format of the `clean_test` variable to make it easier to use. Keep the ordering of the categories, but get rid of the originally-built-in assumption that they are equally spaced. (This is a technicality that you won’t normally need to worry about.)

```
bechdel_13 <- bechdel_13 |>
  mutate(clean_test = factor(as.character(clean_test)),
        clean_test =forcats::fct_relevel(clean_test,
                                         'ok',
                                         'dubious',
                                         'men',
                                         'notalk',
                                         'nowomen'))
```

Your Turn: Analysis

Working with your group, try to use linear regression and the `bechdel_13` dataset to answer:

Are movies with more representation of women less profitable?

Save all your work in an qmd file in your shared RStudio project.

Hints

- The question is intentionally vague. You will have to decide how to measure “representation” and “profitable” (based on available data), and **thoughtfully choose predictor and response variables** to use in your model.

- Before digging in to the data, consider what response you may use and how many predictors you can include, given the size of the dataset.
- Even if you can “afford” to include several predictors in your model, think about whether estimating the effect of one might interfere with estimation or interpretation of another. For example, you probably cannot include *more than one* variable giving results of the Bechdel test (`clean_test`, `binary`) since they are more or less the same information.

Tasks

1. Plan

Write a 1-2 sentence plan before you start working with the data. State which response and predictor(s) you will consider.

2. Exploration

Explore the data using at least one graph that shows your response with at least one predictor.

Talk with your group about patterns you notice.

3. Linear Model

Use `lm()` to fit your planned regression model.

Show the model `summary()` and report the model equation.

4. Conclusions

- Based on the model, try to answer the original question – **Are movies with more representation of women less profitable?** Use the model equation to interpret your findings as best you can.

What problems or questions do you have?

Keep a record of all your work in your shared project, but you do not need to hand anything in.